ONLINE METHODS

Participants. The early-onset IBD discovery case cohort (DC-IBD; **Supplementary Table 1**) consisted of 2,413 Europeans with ancestry cases of IBD (1,636 with Crohn's disease, 724 with ulcerative colitis and 53 with IBD-U) recruited from multiple centers from four geographically discrete countries that met the study's quality-control criteria and were successfully matched with disease-free control subjects from the United States (see **Supplementary Note** for additional details). The Research Ethics Boards of the respective hospitals and other participating centers approved this study, and written informed consent was obtained from all subjects (or their legal guardians).

Genotyping. We performed high-throughput genome-wide SNP genotyping using Illumina Infinium II HumanHap550 BeadChip technology at the Center for Applied Genomics at CHOP as described[18]. After genotyping, we excluded 270 individuals with IBD for whom >2% of genotypes were missing. Control subjects used for this study were also filtered to include those with a genotyping frequency of >98%.

Population stratification. We applied the program STRUCTURE to our quality-controlled data set to exclude 316 individuals with <95% European ancestry on the basis of ancestry informative markers[19]. After these exclusions, our cohort consisted of 2,784 IBD cases, including 1,887 Crohn's disease–only, 835 ulcerative colitis–only and 54 IBD-U cases. Controls were filtered for 95% European ancestry, as determined by STRUCTURE, yielding 7,315 total subjects.

Because of the differing geographical origins between our cases and controls, we performed PCA-based genetic matching (eigenmatching) to minimize intra-European population stratification. Eigenmatching uses singular value decomposition of genotypic data to match cases to their closest controls in the space of *k* principal components. This approach is a variant of a published method[20]; however in contrast to that method, we employed matching as a filtering criterion. Unlike EIGENSTRAT, a common approach to correct for the effects of stratification by adjusting genotype values, eigenmatching removes samples from both cases and controls that are responsible for stratification. We eigenmatched IBD cases and controls through a multistep procedure. First, we computed principal components for our data set by running smartpca in the EIGENSTRAT[21] package on 100,000 random autosomal SNPs. Second, we applied a matching algorithm implemented in MATLAB to the principal components results[21]. This algorithm determines a distance for each case-control pair after mapping each sample to coordinates on the basis of the top *k* eigenvalue-scaled principal components. The algorithm then matches each case to up to *p* nearby controls, keeping only cases that match between *p* and *q* controls within a distance *d* (where *p*, *q*, *d* and *k* are user-specified parameters).

We genetically matched cases in the discovery cohort by using the top seven principal components, matching cases to three genetically related controls, and keeping cases with between one and three controls within a distance of 0.1. We chose $k = 7$ on the basis of the decay plot of the eigenvalues corresponding to the top principal components. We chose 0.1 as a distance threshold *d* after manual optimization minimizing the genomic inflation factor (λ) while maximizing power (that is, number of cases). After matching, we obtained λ values of 1.13, 1.09 and 1.16 for the early-onset IBD, Crohn's disease and ulcerative colitis cohorts, respectively. Our final discovery cohort after matching consisted of 2,413 cases and 6,158 controls, including 1,636 Crohn's disease and 724 ulcerative colitis cases. A summary of the number of subjects who met quality-control and genetic matching criteria for study inclusion is shown in **Supplementary Table 1**. We also performed genetic matching in the replication cohort. Given the smaller number of cases, we employed $k = 7$ principal components to match each case to ten genetically related controls, keeping cases with between one and ten controls within a distance of 0.05. This yielded λ values of 1.16, 1.14 and 1.08 for the IBD, Crohn's disease and ulcerative colitis replication cohorts, respectively. Our final replication cohort RC1 consisted of 482 early-onset IBD cases and 1,696 controls, including 289 Crohn's disease, 120 ulcerative colitis and 73 IBD-U cases.

Replication experiments. We assembled an early-onset IBD replication cohort (RC1) that included cases that were not genetically matched to controls during creation of the discovery cohort and additional cases from the CHOP health system that were obtained through an ongoing collection effort. Additional cases were also genotyped by Illumina Infinium II HumanHap550 BeadChip technology using standard approaches for quality-control filtering (see above). Cases in RC1 were genetically matched to an independent set of European ancestry controls gathered from the CHOP internal collection. After genetic matching, this data set consisted of 482 early-onset IBD cases (289 Crohn's disease, 120 ulcerative colitis and 73 IBD-U) and 1,696 controls. Subsets of data set RC1 corresponding to Crohn's disease and ulcerative colitis subtypes are designated as RC1-CD ($n = 289$) and RC2-UC ($n = 120$).

We used a second early-onset IBD replication cohort (RC2-CD) obtained from the IIBDGC. The IIBDGC replication experiment was based on data from genome-wide scans by the NIDDK[10], WTCCC[11] and a Belgian-French[12] collaboration that have previously been combined to undertake a large-scale meta-analysis[5]. Imputed genotype data (using HapMap II B35 r21 (CEU) as the reference population) was used when directly genotyped data were unavailable for a given cohort (details of the genotyping platforms used in each GWAS have been previously described[5]). In total, 531 early-onset cases (17.1% of the total IIBDGC Crohn's disease cases) and 4,109 population controls were included in the replication effort. Mean age at onset in the IIBDGC early-onset cases was 14.6 yr (s.d. = 3.27 yr). Details of the ascertainment and characterization of the IIBDGC cohort, as well as the quality-control procedures applied to the GWAS data sets, are provided in the original scan and replication publications[10–12,22,23]. Recruitment of study subjects was approved by local and national institutional review boards, and informed consent was obtained from all participants.

Lastly, we followed up select early-onset IBD signals from a meta-analysis of adult-onset Crohn's disease[5], which combined data from three scans totaling 3,230 cases and 4,829 controls.

Association analysis. To detect significantly associated susceptibility alleles, we compared single-marker allele frequencies using χ^2 statistics on SNPs with a minor allele frequency of >1% and with Hardy-Weinberg equilibrium $P > 10^{-5}$. All tests of association were carried out using PLINK[24] and MATLAB with standard criteria for SNP quality-control filtering (see **Supplementary Note**). Given a conservative estimate of less than ~500,000 independent hypotheses, we determined genome-wide significance with a Bonferroni-corrected *P*-value threshold of 1.0×10^{-7}. We also examined nominal signals below a *P*-value threshold of 1×10^{-6}. We excluded 'loner' signals whose significance level was discordant with that of adjacent SNPs in their LD or genomic neighborhood. SNP coordinates were obtained from the National Center for Biotechnology Information (NCBI) Build 36 and LD information was obtained using HapMap II B36 r27 (CEU). We used PLINK and MATLAB to determine nominal *P* values, ORs and confidence intervals for ORs for all SNPs tested. We compared our association results with SNPs in published data sets by using either the exact SNP or the best LD surrogate ($r^2 > 0.2$) found on our scan.

We combined data from multiple scans for meta-analysis by computing a *Z* score at each SNP. In brief, *P* values in each scan *i* were transformed via the inverse normal cumulative distribution function into a *Z* score z_i, with signs of the score indicating direction of effect relative to the minor allele, that is, positive *Z* scores for risk-conferring variants (OR > 1) and negative *Z* scores for protective variants (OR < 1). *Z* scores z_i from individual studies were summed into a z_{meta} value using weights w_i for each scan computed as $\text{sqrt}(N_i/N)$ where N_i is the number of individuals in study *i* and *N* is the total number of individuals across all studies. The combined *Z* score z_{meta} was transformed into a *P* value via the normal cumulative distribution function. Z scores were computed for 299,238, 489,951 and 299,238 SNPs in the Crohn's disease, ulcerative colitis and IBD meta-analyses, respectively. All meta-analysis computations were implemented and performed using MATLAB.

Gene expression analysis. We examined allele-specific effects on gene expression for significantly associating loci by assaying total RNA in genotyped LCLs. We also compared gene expression between colonic biopsy specimens obtained from early-onset IBD cases and normal controls to detect disease-specific gene expression differences. We evaluated allele-specific effects on the expression of genes *IL27* and *EIF3C* for the rs1968752 variant on 16p11 and genes *HORMAD2* and *LIF* for the rs2412973 variant on 21q22 (see

Supplementary Note). In brief, RNA was isolated from HapMap-CEU population samples using Trizol (Invitrogen). Real-time RT PCR was performed on a Bio-Rad iCycler System using SYBR Green detection (Bio-Rad). cDNA template was made from 2 μg of total RNA by using the Invitrogen cDNA Synthesis kit. Primer sequences were designed using Integrated DNA Technologies (IDT). β-Actin was used as a control. Each reaction was carried out in triplicate wells on one plate. Fold change between the A/A and C/C genotype was calculated with the comparative C_T method. Results were normalized to β-actin for cDNA quantification. Data were analyzed by analysis of variance (ANOVA). We also examined allele-specific changes on gene expression in a publicly available LCL database[14].

We examined colonic expression of selected candidate genes located in the LD blocks of our most significant signals. Gene expression was assayed in individual colonic biopsy specimens from subjects with early-onset Crohn's disease ($n = 30$) and early-onset ulcerative colitis ($n = 10$), and from healthy controls ($n = 11$). Individuals were aged from 5 to 18 years at time of biopsy with a median age of 13 years. Inflammation was quantified in colon biopsies by using the Crohn's Disease Histological Index of Severity, and ranged from grade 1 to 12 with a median of 3.5. Of the 30 individuals with Crohn's disease, 16 were receiving one or more of the following medications at time of biopsy: 5-ASA, cytoxan, 6-mercaptopurine and methotrexate. After informed consent, colonic biopsies were obtained from subjects with Crohn's disease and ulcerative colitis, and healthy controls. All of the biopsies for IBD cases and healthy controls were obtained from the ascending colon, with the exception of one subject with ulcerative colitis whose biopsy was obtained from the rectum. Colon biopsies were immediately placed in RNAlater stabilization reagent (Qiagen, Germany) at 4 °C. Total RNA was isolated by an RNeasy Plus Mini Kit (Qiagen) and stored at –80 °C. Samples were then submitted to the CCHMC Digestive Health Center Microarray Core where the quality and concentration of RNA were measured by the Agilent Bioanalyser 2100 (Hewlett Packard) using an RNA 6000 Nano Assay to confirm a 28S/18S ratio of 1.6-2.0. We amplified 100 ng of total RNA by using a Target 1-round Aminoallyl-aRNA Amplification Kit 101 (Epicentre, WI). The biotinylated cRNA was hybridized to Affymetrix GeneChip Human Genome HG-U133 Plus 2.0 arrays, containing probes for 22,634 genes. The images were captured by an Affymetrix Genechip Scanner 3000. The complete data set is available at the NCBI Gene Expression Omnibus. Of note, this data set includes two additional expression profiles for two individuals with early-onset Crohn's disease who were biopsied both pre- and posttherapy. The posttherapy samples for these individuals were not used in analyses for this paper (that is, we used gene expression values only from pretherapy samples). Significance of gene expression changes was assessed by one-way ANOVA with Tukey-Kramer multiple comparison correction and P-value thresholds of $P < 0.05$, $P < 0.001$ and $P < 0.0001$. Genes demonstrating significant expression differences were further evaluated by one-way analysis of covariance (ANCOVA) with Tukey-Kramer multiple comparison correction and the same P-value thresholds, applying histological inflammation score and number of concurrent medication as covariates to assess the impact of these clinicopathological factors on the expression results. We also determined allele-specific colonic gene expression in colonic tissue for our most significantly associated SNPs. For this, we used a subset of IBD cases ($n = 13$) for which both gene expression was measured and genome-wide genotyping was done (as part of the current study). We determined allele-specific expression by one-way ANOVA with a significance threshold of 0.05. All gene expression analyses and statistical tests were implemented and performed in MATLAB.

URL. GEO, http://www.ncbi.nlm.nih.gov/geo.

18. Hakonarson, H. *et al.* A genome-wide association study identifies *KIAA0350* as a type 1 diabetes gene. *Nature* **448**, 591–594 (2007).
19. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
20. Luca, D. *et al.* On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. *Am. J. Hum. Genet.* **82**, 453–463 (2008).
21. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
22. Duerr, R.H. *et al.* A genome-wide association study identifies *IL23R* as an inflammatory bowel disease gene. *Science* **314**, 1461–1463 (2006).
23. Parkes, M. *et al.* Sequence variants in the autophagy gene *IRGM* and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat. Genet.* **39**, 830–832 (2007).
24. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

doi:10.1038/ng.489